https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

Predicting Gas-Particle Partitioning Coefficients of Atmospheric Molecules with Machine Learning

Emma Lumiaro[1], Milica Todorović[1], Theo Kurten[2], Hanna Vehkamäki[3], and Patrick Rinke[1]

[1]Department of Applied Physics, Aalto University, P.O. Box 11100, 00076 Aalto, Espoo, Finland
[2]Department of Chemistry, Faculty of Science, PO Box 55, FI-00014 University of Helsinki, Finland
[3]Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, PO Box 64, FI-00014 University of Helsinki, Finland

Correspondence: Patrick Rinke (patrick.rinke@aalto.fi)

Abstract. The formation, properties and lifetime of secondary organic aerosols in the atmosphere are largely determined by gas-particle partitioning coefficients of the participating organic vapours. Since these coefficients are often difficult to measure and to compute, we developed a machine learning model to predict them given molecular structure as input. Our data-driven approach is based on the dataset by Wang et al. (Atmos. Chem. Phys., 17, 7529 (2017)), who computed the partitioning coefficients and saturation vapour pressures of 3414 atmospheric oxidation products from the master chemical mechanism using the COSMOtherm program. We trained a kernel ridge regression (KRR) machine learning model on the saturation vapour pressure ($\mathrm{P_{sat}}$), and on two equilibrium partitioning coefficients: between a water-insoluble organic matter phase and the gas phase ($\mathrm{K_{WIOM/G}}$), and between an infinitely dilute solution with pure water and the gas phase ($\mathrm{K_{W/G}}$). For the input representation of the atomic structure of each organic molecule to the machine, we tested different descriptors. We find that the many-body tensor representation (MBTR) works best for our application, but the topological fingerprint (TopFP) approach is almost as good, and is significantly more cost effective. Our best machine learning model (KRR with a Gaussian kernel + MBTR) predicts $\mathrm{P_{sat}}$ and $\mathrm{K_{WIOM/G}}$ to within 0.3 logarithmic units and $\mathrm{K_{W/G}}$ to within 0.4 logarithmic units of the original COSMOtherm calculations. This is equal or better than the typical accuracy of COSMOtherm predictions compared to experimental data (where available). We then applied our machine learning model to a dataset of 35,383 molecules that we generated based on a carbon 10 backbone functionalized with 0 to 6 carboxyl, carbonyl or hydroxyl groups to evaluate its performance for polyfunctional compounds with potentially low $\mathrm{P_{sat}}$. The resulting saturation vapor pressure and partitioning coefficient distributions were physico-chemically reasonable, and the volatility predictions for the most highly oxidized compounds were in qualitative agreement with experimentally inferred volatilities of atmospheric oxidation products with similar elemental composition.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

1 Introduction

Aerosols in the atmosphere are fine solid or liquid particles (or droplets) suspended in air. They scatter and absorb solar radiation and form cloud droplets in the atmosphere, affect visibility and human health and are responsible for large uncertainties in the study of climate change. Most aerosol particles are secondary organic aerosols (SOAs) that are formed by oxidation of volatile organic compounds (VOCs), which are in turn emitted into the atmosphere for example from plants or traffic. Some of the oxidation products have volatilities low enough to condense. The formation, growth and lifetime of SOAs is governed largely by the concentrations, saturation vapour pressures (P_{sat}) and equilibrium partitioning coefficients of the participating vapours. While real atmospheric aerosol particles are extremely complex mixtures of many different organic and inorganic compounds (Elm et al., 2020), partitioning of organic vapours is by necessity usually modelled in terms of a few representative parameters. These include the saturation vapour pressure, describing the interaction of a compound with itself, and various partitioning coefficients (K) for the interaction of the compound with representative other species. Typical partitioning coefficients in chemistry include ($\mathrm{K_{W/G}}$) for the partitioning between the gas phase and pure water (i.e. an infinitely dilute solution of the compound), and ($\mathrm{K_{O/W}}$) for the partitioning between octanol and water solutions [1]. For organic aerosols, the partitioning coefficient between the gas phase and a model water-insoluble organic matter phase (WIOM; $\mathrm{K_{WIOM/G}}$) is more appropriate than ($\mathrm{K_{O/G}}$).

Unfortunately, experimental measurements of these partitioning coefficients are challenging, especially for multifunctional low-volatility compounds most relevant to SOA formation. Little experimental data is thus available for the atmospherically most interesting organic vapour species. For relatively simple organic compounds, efficient empirical parametrizations have been developed to predict their condensation-relevant properties. These include poly-parameter linear free-energy relationships (ppLFERs) (Goss and Schwarzenbach, 2001; Goss, 2004, 2006), the GROup contribution Method for Henry's law Estimate (GROMHE) (Raventos-Duran et al., 2010), and SPARC Performs Automated Reasoning in Chemistry (SPARC) (Hilal et al., 2008), SIMPOL (Pankow and Asher, 2008), EVAPORATION (Compernolle et al., 2011), and Nannoolal (Nannoolal et al., 2008). Many of these parameterisations are available in a user-friendly format on the UManSysProp website (Topping et al., 2016). However, due to the limitations in the available experimental datasets on which they are based, the accuracy of such approaches typically degrades significantly once the compound contains more than three or four functional groups (Valorso et al., 2011).

Approaches based on quantum chemistry such as COSMO-RS (COnductor-like Screening MOdel for Real Solvents, Klamt and Eckert (2000, 2003); Eckert and Klamt (2002)), implemented for example in the COSMOtherm program, can calculate saturation vapour pressures and partitioning coefficients also for complex polyfunctional compounds, albeit only with order-of-magnitude accuracy. However, for many applications even this is extremely useful. For example, in the context of new-particle formation (often called nucleation) it is useful to know, if the saturation vapour pressure of an organic compound is lower than about 10^{-12} kPa, because then it could condense irreversibly onto preexisting nanometer-sized cluster. An even lower P_{sat} would be required for the vapour to form completely new particles. This illustrates the challenge in performing experiments

[1] The gas-octanol partitioning coefficient ($\mathrm{K_{O/G}}$) can then be obtained from these by division.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021


Atmospheric Chemistry and Physics Discussions

on SOA-relevant species: a compound with a saturation vapour pressure of e.g. 10^{-8} kPa at room temperature would be considered non-volatile in terms of most available measurement methods – yet its volatility is far too high to allow nucleation in the atmosphere.

COSMO-RS/COSMOtherm calculations are based on density functional theory (DFT). In the context of quantum chemistry they are therefore considered computationally tractable compared to high-level methods such as coupled cluster theory. Nevertheless, the application of COSMO-RS to complex polyfunctional organic molecules still entails a significant computational effort, especially due to the conformational complexity of these species that need to be taken into account appropriately. Overall, there could be up to $10^4 - 10^7$ different organic compounds in the atmosphere (not even counting most oxidation intermediates), which makes the computation of saturation vapour pressures and partitioning coefficients a daunting task (Shrivastava et al., 2019; Ye et al., 2016).

Here, we take a different approach compared to previous parametrization studies, and consider a data-science perspective (Himanen et al., 2019). Instead of assuming chemical or physical relations, we let the data speak for itself. We develop and train a machine learning model to extract patterns from available data and predict saturation vapour pressures as well as partitioning coefficients.

Machine learning has only recently spread into atmospheric science (Cervone et al., 2008; Toms et al., 2018; Barnes et al., 2019; Nourani et al., 2019; Huntingford et al., 2019; Masuda et al., 2019). Prominent applications include the identification of forced climate patterns (Barnes et al., 2019), precipitation prediction (Nourani et al., 2019), climate analysis (Huntingford et al., 2019), pattern discovery (Toms et al., 2018), risk assessment of atmospheric emissions (Cervone et al., 2008), and the estimation of cloud optical thicknesses (Masuda et al., 2019). In molecular and materials science, machine learning is more established and now frequently complements theoretical or experimental methods (Müller et al., 2016; Ma et al., 2015; Shandiz and Gauvin, 2016; Gómez-Bombarelli et al., 2016; Bartók et al., 2017; Rupp et al., 2018; Goldsmith et al., 2018; Meyer et al., 2018; Zunger, 2018; Gu et al., 2019; Schmidt et al., 2019; Jensen et al.; Coley et al.). Here we build on our experience in atomistic, molecular machine learning (Ghosh et al., 2019; Todorović et al., 2019; Stuke et al., 2019; Himanen et al., 2020) to train a regression model that maps molecular structures onto saturation vapour pressures and partitioning coefficients. Once trained, the machine learning model can make saturation vapour pressure and partitioning predictions at COSMOtherm accuracy for hundreds of thousands of new molecules at no further computational cost. When experimental training data becomes available, the machine learning model could easily be extended to encompass predictions for experimental pressures and coefficients.

Due to the above-mentioned lack of comprehensive experimental databases for saturation vapour pressures or gas-liquid partioning coefficient of polyfunctional atmospherically relevant molecules, our machine-learning model is based on the computational data by Wang et al. (2017). They computed the partitioning coefficients and saturation vapour pressures for 3414 atmospheric secondary oxidation products, obtained from the Master Chemical Mechanism (Jenkin et al., 1997; Saunders et al., 2003), using a combination of quantum chemistry and statistical thermodynamics as implemented in the COSMOtherm approach (Klamt and Eckert, 2000).

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

We transform the molecular structures in Wang's dataset into atomistic descriptors more suitable for machine learning than the atomic coordinates or the commonly used simplified molecular-input line-entry system (SMILES) strings. Optimal descriptor choices have been the subject of increased research in recent years (Langer et al., 2020; Rossi and Cumby, 2020; Himanen et al., 2020). We here test several descriptor choices: the many body tensor representation (Huo and Rupp, 2017), the Coulomb matrix (Rupp et al., 2012), the Molecular ACCess System (MACCS) structural key (Durant et al., 2002), a topological fingerprint developed by RDkit (Landrum et al., 2006) based on the daylight fingerprint (James et al., 1995) and the Morgan fingerprint (Morgan, 1965).

Our work addresses the following objectives: 1) With view to future machine learning applications in atmospheric science, we assess the predictive capability of different structural descriptors for machine learning the chosen target properties. 2) We quantify the predictive power of our KRR machine learning model for Wang's dataset to ascertain, if the dataset size is sufficient for accurate machine learning predictions. 3) We then apply our validated machine learning model to a new molecular dataset to gain chemical insight into SOA condensation processes.

The paper is organized as follows. We describe our machine learning methodology in section 2, then present the machine learning results in section 3. Section 4 demonstrates how we employed the trained model for fast prediction of molecular properties. We discuss our findings and present a summary in section 5.

2 Methods

Figure 1. Schematic representation of our machine learning workflow

Our machine learning approach has five components as illustrated in Fig. 1. We start off with the raw data, which we present and analyse in section 2.1. The raw data is then transformed into a suitable representation for machine learning (step 2). We introduce five different representations in section 2.2, which we test in our machine learning model (cf section 3). Next we choose our machine learning method. Here we use kernel ridge regression (KRR), which is introduced in section 2.3. We analyse the learning success of our machine learning approach in step 4. The results of this process are shown in section 3. In this step we also make adjustments to the representation and the parameters of the model to improve the learning. Finally, we use the best machine learning model to make predictions as shown in section 4.

2.1 Dataset

In this work we are interested in the the equilibrium partitioning coefficients of a molecule between a water-insoluble organic matter (WIOM) phase and gas phase ($K_{WIOM/G}$) as well as gas phase and infinitely dilute water solution. These coefficients

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

are defined as

$$K_{\mathrm{WIOM/G}} = \frac{C_{\mathrm{WIOM}}}{C_{\mathrm{G}}} \tag{1}$$

$$K_{\mathrm{W/G}} = \frac{C_{\mathrm{W}}}{C_{\mathrm{G}}}, \tag{2}$$

where C_{WIOM}, C_{W}, and C_{G} are the equilibrium concentrations of the molecule in the WIOM, water, and gas phase, respectively, at the limit of infinite dilution. In the framework of COSMOtherm calculations, gas-liquid partitioning coefficients can be converted into saturation vapor pressures, or vice versa, using the activity coefficients γ_{W} or γ_{WIOM} in the corresponding liquid (which can also be computed by COSMOtherm). Specifically, if for example $K_{\mathrm{W/G}}$ is expressed in units of $m^3 g^{-1}$, then $P_{\mathrm{sat}} = \frac{RT}{M \gamma_{\mathrm{W}} K_{\mathrm{W}}}$, where R is the gas constant, T the temperature, M the molar mass of the compound and K_{W} and γ_{W} are the partitioning and activity coefficients in water (Arp and Goss, 2009). We caution, however, that many different conventions exist e.g. for the dimensions of the partitioning coefficients, as well as the reference states for activity coefficients – the relation given above applies only to the particular conventions used by COSMOtherm.

Wang et al. (2017) used the conductor-like screening model for real solvents (COSMO-RS) theory (Klamt and Eckert, 2000) implemented in COSMOtherm to calculate the two partitioning coefficients $\mathrm{K_{WIOM/G}}$[2] and $\mathrm{K_{W/G}}$ for 3414 molecules. These molecules were generated from 143 parent volatile organic compounds with the Master Chemical Mechanism (MCM) (Jenkin et al., 1997; Saunders et al., 2003) through photolysis and reactions with ozone, hydroxide radicals and nitrate radicals.

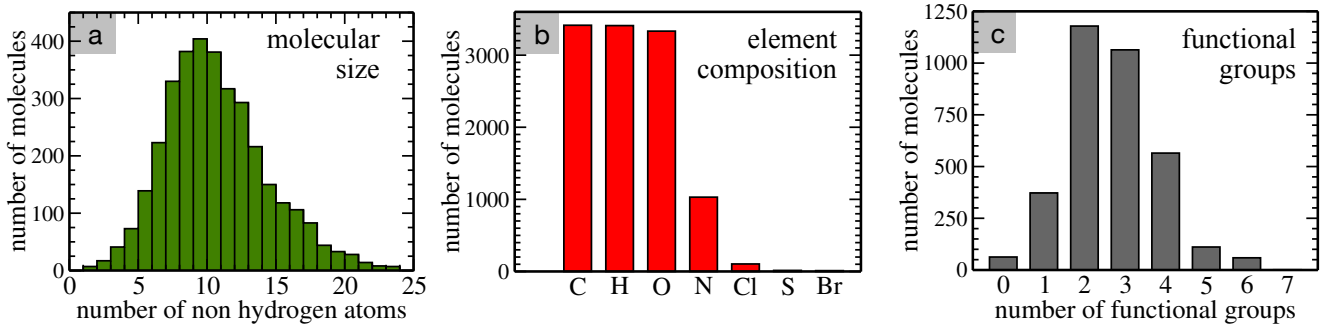

Figure 2. Dataset statistics: Panel a) shows the size distribution (in terms of the number of non-hydrogen atoms) of all 3414 molecules in the dataset. Panel b) illustrates how many molecules contain each of the chemical species and panel c) depicts the functional group distribution.

Here, we analyse the composition of the publicly available dataset by Wang *et al.* in preparation for machine learning. Figure 2 illustrates key dataset statistics. Panel a) shows the size distribution of molecules as measured in the number of non-hydrogen atoms. The 3414 non-radical species obtained from MGM range in size from 4 to 48 atoms, which translates into 2 to 24 non-hydrogen atoms per molecule. The distribution peaks at 10 non-hydrogen atoms and is skewed towards larger

[2]As a model WIOM phase Wang *et al.* used a compound originally suggested by Kalberer et al. (2004) as a representative secondary organic aerosol constituent. The IUPAC name for the compound in question, with elemental composition $C_{14}H_{16}O_5$, is 1-(5-(3,5-dimethylphenyl)dihydro-[1,3]dioxolo[4,5-d][1,3]dioxol-2-yl)ethan-1-one.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

molecules. Panel b) illustrates how many molecules contain at least one atom of the indicated element. All molecules contain carbon (100% C), 3410 contain hydrogen (H; 99.88%) and 3333 also oxygen (O; 97.63%). Nitrogen (N) is the next most abundant element (30.17%) followed by chlorine (Cl; 3.05%), sulphur (S; 0.44%) and bromide (Br; 0.32%). Lastly, panel c) presents the distribution of functional groups. It peaks at 2 (34%) to 3 (31%) functional groups per molecule, with relatively few molecules having 0 (2%), 5 (3%) or 6 (2%) functional groups. The percentages for 1 and 4 functional groups are 11% and 17%, respectively.

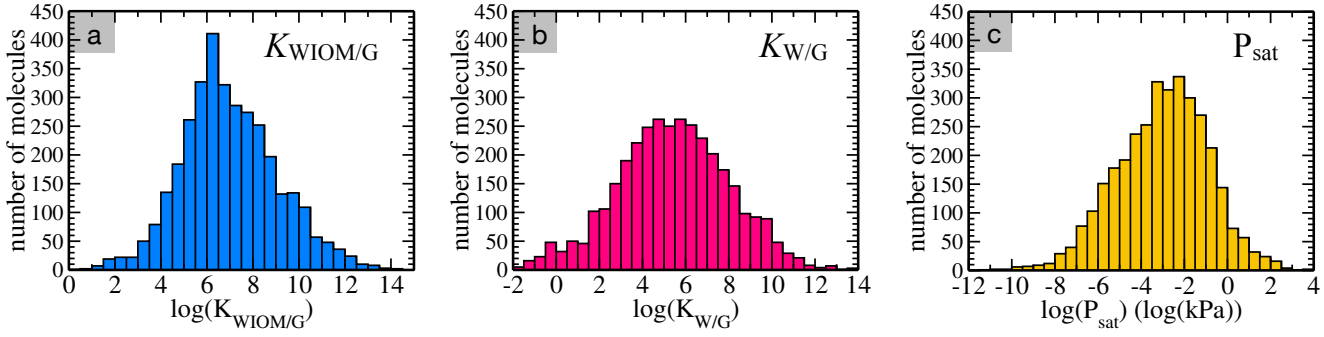

Figure 3. Dataset statistics: distributions of equilibrium partitioning coefficients a) $K_{WIOM/G}$, b) $K_{W/G}$ and c) the saturation vapour pressure P_{sat} for all 3414 molecules in the dataset.

Figure 3 shows the distribution of the target properties $K_{WIOM/G}$, $K_{W/G}$ and P_{sat} in Wang's dataset on a logarithmic scale. The equilibrium partitioning coefficient $K_{WIOM/G}$ distribution is skewed slightly towards larger coefficients, in contrast to the saturation vapour pressure P_{sat} distribution with an asymmetry towards molecules with lower pressures. All three target properties cover approximately 15 logarithmic units and are approximately Gaussian distributed. Such peaked distributions are often not ideal for machine learning since they over-represent molecules near the peak of the distribution and under-represent molecules at their edges. The data peak does supply enough similarity to ensure good quality learning, but properties of the under-represented molecular types might be harder to learn.

Wang's dataset of 3414 molecules is relatively small for machine learning, which often requires hundreds of thousands to millions of training samples (Pyzer-Knapp et al.; Smith et al., 2017; Stuke et al., 2019; Ghosh et al., 2019). A slightly larger set of Henry's law constants, which are related to $K_{W/G}$, were reported by Sander (2015) for 4632 organic species. Sander's database is a collection of 17350 Henry's law constant values collected from 689 references and therefore not as internally consistent as Wang's dataset. We are not aware of a larger dataset that reports partitioning coefficients. For this reason, we rely exclusively on Wang's dataset and show that we can develop machine learning methods that are just as accurate as the underlying calculations and thus suitable for predictions.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021


Atmospheric Chemistry and Physics Discussions

2.2 Representations

The molecular representation for machine learning should fulfil certain requirements. It should be invariant with respect to translation and rotation of the molecule and permutations of atomic indices. Furthermore, it should be continuous, unique, compact and efficient to compute (Faber et al., 2015; Huo and Rupp, 2017; Langer et al., 2020; Himanen et al., 2020).

In this work we employ two classes of representations for the molecular structure, also known as descriptors: *physical* and *cheminformatics* descriptors. *Physical descriptors* encode physical distances and angles between atoms in the material or molecule. Such descriptors generally exhibit good performance for many different system types. Meanwhile, decades of research in *cheminformatics* have produced topological descriptors that encode the qualitative aspects of molecules in a compact representation. These descriptors are typically bitvectors, in which molecular features are encoded (hashed) into binary fingerprints, which are joined into long binary vectors. In this work, we use two physical descriptors, the Coulomb Matrix and the many-body tensor, and three cheminformatics descriptors: the MACCS structural key, the topological fingerprint and the Morgan fingerprint.

In Wang's dataset the molecular structure is encoded in SMILES (Simplified Molecular Input Line Entry Specification) strings. We convert these SMILES strings into structural descriptors using Open Babel (O'Boyle et al., 2011) and the DScribe library (Himanen et al., 2020) or into cheminformatics descriptors using RDkit (Landrum et al., 2006).

2.2.1 Coulomb Matrix

The Coulomb matrix (CM) descriptor is inspired by an electrostatic representation of a molecule (Rupp et al., 2012). It encodes the cartesian coordinates of a molecule in a simple matrix of the form

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{\|\mathbf{R}_i - \mathbf{R}_j\|} & \text{if } i \neq j \end{cases} \tag{3}$$

where $\mathbf{R}_i$ is the coordinate of atom i with atomic charge Z_i. The diagonal provides element-specific information. The coefficient and the exponent have been fitted to the total energies of isolated atoms (Rupp et al., 2012). Off-diagonal elements encode inverse distances between the atoms of the molecule by means of a Coulomb-repulsion-like term.

The dimension of the Coulomb matrix is chosen to fit the largest molecule in the data set, i.e. it corresponds to the number of atoms of the largest molecule. The "empty" rows of Coulomb matrices for smaller molecules are padded with zeroes. Invariance with respect to the permutation of atoms in the molecule is enforced by simultaneously sorting rows and columns of each Coulomb matrix in descending order according to their ℓ^2-norms. An example of a Coulomb matrix for 2-hydroxy-2-methylpropanoic acid is shown in Fig. 4b.

The CM is easily understandable, simple and relatively small as a descriptor. However, it performs best with Laplacian kernels in the machine-learning model (see Section 2.3), while other descriptors work better with the more standard choice of a Gaussian kernel.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

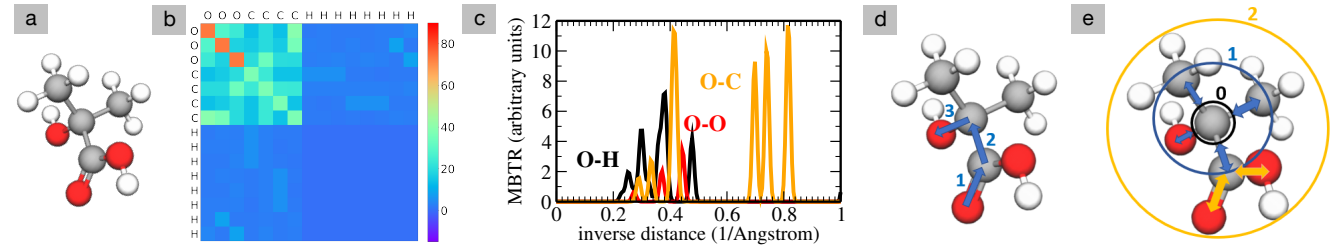

Figure 4. Pictorial overview over descriptors used in this work: a) ball and stick model of 2-hydroxy-2-methylpropanoic acid, b) corresponding Coulomb matrix (CM), c) the O-H, O-O and O-C inverse distance entries of the many-body tensor representation (MBTR), d) topological fingerprint (TopFP) depiction of a path with length three, and e) Morgan circular fingerprint with radius 0 (black), radius 1 (blue) and radius 2 (orange).

2.2.2 Many-body tensor representation

The many-body tensor representation (MBTR) follows the Coulomb matrix philosophy of encoding the internal coordinates of a molecule. We will here describe the MBTR only qualitatively. Detailed equations can be found in the original publication (Huo and Rupp, 2017), our previous work (Himanen et al., 2020; Stuke et al., 2020) or Appendix A.

Unlike the Coulomb matrix, the many-body tensor is continuous and it distinguishes between different types of internal coordinates. At many-body level 1, the MBTR records the presence of all atomic species in a molecule by placing a Gaussian at the atomic number on an axis from 1 to the number of elements in the periodic table. The weight of the Gaussian is equal to the number of times the species is present in the molecule. At many-body level 2, inverse distances between every pair of atoms (bonded and non-bonded) are recorded in the same fashion. Many-body level 3 adds angular information between any triple of atoms. Higher levels (e.g. dihedral angles) would in principle be straightforward to add, but are not implemented in the current MBTR versions (Huo and Rupp, 2017; Himanen et al., 2020). Figure 4c shows selected MBTR elements for 2-hydroxy-2-methylpropanoic acid.

The MBTR is a continuous descriptor, which is advantageous for machine learning. However, MBTR is by far the largest descriptor out of the five we tested, and this can pose restrictions on memory and computational cost. Furthermore, the MBTR is more difficult to interpret than the CM.

2.2.3 MACCS Structural Key

The Molecular ACCess System (MACCS) structural key is a dictionary-based descriptor (Durant et al., 2002). It is represented as a bitvector of Boolean values that encode answers to a set of predefined questions. The MACCS structural key we used is a 166 bit long set of answers to 166 questions such as "Is there an S-S bond" or "Does it contain Iodine?" (Landrum et al., 2006; James et al., 1995).

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021


Atmospheric Chemistry and Physics Discussions

MACCS is the smallest out of the five descriptors and extremely fast to use. Its accuracy critically depends on how well the 166 questions encapsulate the chemical detail of the molecules. Is it likely to reach a moderate accuracy with low computational cost and memory usage and could be beneficial for fast testing of a machine learning model.

2.2.4 Topological Fingerprint

The topological fingerprint (TopFP) is RDKit's original fingerprint (Landrum et al., 2006) inspired by the Daylight fingerprint (James et al., 1995). TopFP hashes the atomic structure. This means that the structure is divided into smaller substructures and each substructure is converted into a unique binary ID called a hash. The hashes are concatenated into a long bitvector representing the entire molecule. In TopFP, the molecule is hashed along topological paths, or along bonds, as illustrated in Figure 4d. The path starts from one atom in a molecule and travels along bonds until k bond lengths have been traversed, and that completes a hash. The length of the bitvector, maximum and minimum possible path lengths k_{max} and k_{min} and the length of one hash can be optimized.

Topology is an informative molecular feature. We therefore expect TopFP to balance good accuracy with reasonable computational cost. However, this binary fingerprint is difficult to visualize and analyse for chemical insight.

2.2.5 Morgan Fingerprint

The Morgan fingerprint is also a bit-vector constructed by hashing the molecular structure. In contrast to the Topological fingerprint, the Morgan fingerprint is hashed along circular or spherical paths around the central atom as illustrated in Figure 4e. Each substructure for a hash is constructed by first numbering the atoms in a molecule with unique integers by applying the Morgan algorithm. Each uniquely numbered atom then becomes a cluster center, around which we iteratively increase a spherical radius to include the neighbouring bonded atoms (Rogers and Hahn, 2010). Each radius increment extends the neighbour list by another molecular bond. The "circular" substructures found by the algorithm described above, excluding duplicates, are then hashed into a fingerprint (James et al., 1995; Landrum et al., 2006). The length of the fingerprint and the maximum radius can be optimized.

The Morgan fingerprint is quite similar to the TopFP in size and type of information encoded, so we expect similar performance. It also does not lend itself to easy chemical interpretation.

2.3 Machine Learning Method

2.3.1 Kernel Ridge Regression

In this work, we apply the kernel ridge regression (KRR) machine learning method. KRR is an example of supervised learning, in which the machine learning model is trained on pairs of input (x) and target (f) data. The trained model then predicts target values for previously unseen inputs. In this work, the input x are the molecular descriptors CM and MBTR as well as the MACCS, TopFP and Morgan fingerprints. The targets are scalar values for the equilibrium partitioning coefficients and saturation vapour pressures.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021


Atmospheric Chemistry and Physics Discussions

KRR is based on Ridge Regression, in which a penalty for overfitting is added to an ordinary least squares fit (Friedman et al., 2001). In KRR, unlike Ridge regression, a nonlinear kernel is applied. This maps the molecular structure to our target properties in a high dimensional space (Stuke et al., 2019; Rupp, 2015).

The target values f are a linear expansion in kernel elements

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x), \tag{4}$$

where the sum runs over all training molecules. In this work, we use two different kernels, the Gaussian kernel

$$k_G(x, x') = e^{-\gamma \|x - x'\|_2^2} \tag{5}$$

and the Laplacian kernel

$$k_L(x, x') = e^{-\gamma \|x - x'\|_1}. \tag{6}$$

The kernel width γ is a hyperparameter of the KRR model.

The regression coefficients α_i can be solved by minimizing the error

$$\min_{\alpha} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}, \tag{7}$$

where y_i are reference target values for molecules in the training data. The second term is the regularization term, whose size is controlled by the hyperparameter λ. $\mathbf{K}$ is the kernel matrix of training inputs $k(x_i, x_j)$.

This minimization problem can be solved analytically for the expansion coefficients α_i

$$\boldsymbol{\alpha} = (\mathbf{K} - \lambda \mathbf{I})^{-1} \mathbf{y} \tag{8}$$

The hyperparameters γ and λ need to be optimised separately.

We implemented KRR in Python using *scikit-learn* (Pedregosa et al., 2011). Our implementation has been described in Ref. Stuke et al. (2019, 2020)

2.3.2 Computational Execution

Data used for supervised machine learning is typically divided into two sets, a large training set and a small test set. Both sets consists of input vectors and corresponding target properties. The training set is used to train the KRR model, while the test set molecules are unseen by the trained model. The test set thus quantifies the model performance. At the outset, we separate a test set of 414 molecules. From the remaining molecules, we choose six different training sets of size 500, 1000, 1500, 2000, 2500 and 3000, so that a smaller training size is always a subset of the larger one. Training the model on a sequence of such training sets allows us to compute a *learning curve*, which facilitates the assessment of learning success with increasing training data size. We quantify the accuracy of our KRR model by computing the mean absolute error (MAE) for the test set.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

To get statistically meaningful results, we repeat the training procedure 10 times. In each run, we shuffle the dataset before selecting the training and test sets so that the KRR model is trained and tested on different data each time. Each point on the learning curves is computed as the average over 10 results, and the spread serves as the standard deviation of the datapoint.

Model training proceeds by computing the KRR regression coefficients α_i, obtained by minimizing equation 7. KRR hyperparameters γ and λ are typically optimized via grid search, and average optimal solutions are obtained by cross-validating the procedure. In cross-validation we split off a validation set from the training data before training the KRR model. KRR is then trained for all possible combinations of discretised hyperparameters (grid search) and evaluated on the validation set. This is done several times, so that the molecules in the validation set are changed each time. Then the hyperparameter combination with minimum average cross-validation error is chosen. Our implementation of cross-validated grid search is also based on Scikit-learn (Pedregosa et al., 2011).

Table 1. All the hyperparameters that were optimized.

	Hyperparameters	Optimized Values
KRR	width of the kernel γ, regularization parameter λ	descriptor-dependent
MBTR	broadening parameters σ_2, σ_3; weighting parameters w_2, w_3	0.0075, 0.1; 1.2, 0.8
TopFP	vector length; maximum path length k_{max} ; bits per hash	8192; 8; 16
Morgan	vector length; radius;	2048; 2

Table 1 summarises all the hyperparameters optimised in this study, those for KRR and the molecular descriptors, and their optimal values. In grid search, we varied both γ and λ by ten values between 10^{-1} and 10^{10}. In addition, we used two different kernels, Laplacian and Gaussian. We compared the performance of the two kernels for the average of 5 runs for each training size and the most optimal kernel was chosen. In cases in which both kernels performed equally well, e.g., for the fingerprints, we chose the Gaussian kernel for its lower computational cost.

MBTR hyperparameters and TopFP hyperparameters were optimized by grid search for several training set sizes (MBTR for sizes 500, 1500 and 3000 and TopFP for size 1000 and 1500) and the average of two runs for each training size was taken. We did not extend the descriptor hyperparameter search to larger training set sizes, since we found that the hyperparameters were insensitive to the training set size. The MBTR weighting parameters were optimized in 8 steps between 0 (no weighting) and 1.4, and the broadening parameters in 6 steps between 10^{-1} and 10^{-6}. The length of TopFP was varied between 1024 and 8192 (size can varied by 2^n). The range for the maximum path length extended from 5 to 11 and the bits per hash were varied between 3 and 16.

3 Results

In Figure 5 we present the learning curves for our objectives $K_{\mathrm{WIOM/G}}$, $K_{\mathrm{W/G}}$ and P_{sat}. Shown is the mean average error (MAE) as a function of the training set size for all three target properties and for all five molecular descriptors. As expected,

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

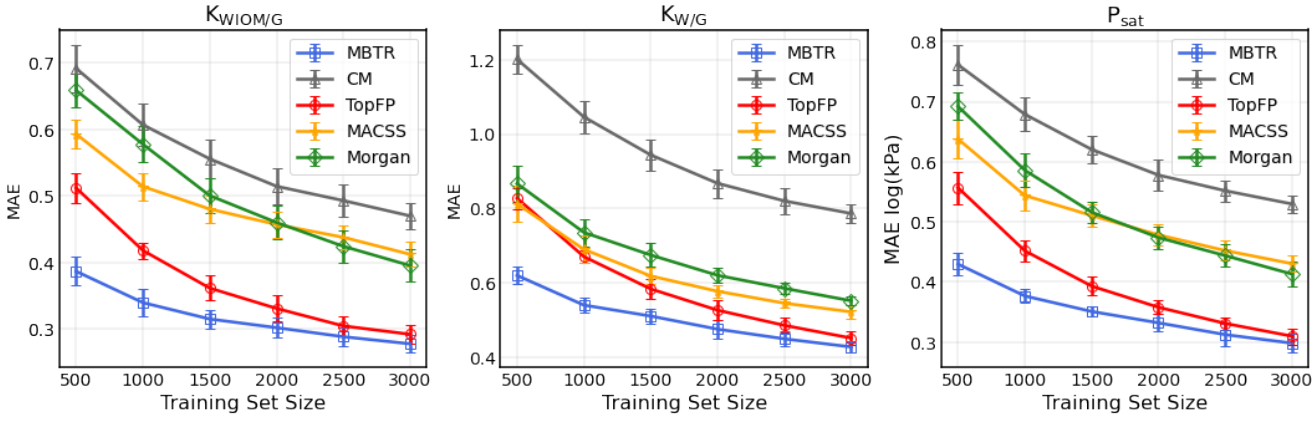

Figure 5. The learning curves for equilibrium partitioning coefficients $K_{\mathrm{WIOM/G}}$, $K_{\mathrm{W/G}}$ and saturation vapour pressure P_{sat} for predictions made with all five descriptors.

the MAE decreases as the training size increases. For all target properties, the lowest errors are achieved with MBTR and the worst performing descriptor is CM. TopFP approaches the accuracy of MBTR as the training size increases and appears likely to outperform MBTR beyond the largest training size of 3000 molecules.

Table 2 summarises the average MAEs and their standard deviations for the best-trained KRR model (training size of 3000 with MBTR descriptor). The highest accuracy is obtained for partitioning coefficient $K_{\mathrm{WIOM/G}}$, with a mean average error of 0.278, i.e. only 1.9% of the entire $K_{\mathrm{WIOM/G}}$ range. The second best accuracy is obtained for saturation vapour pressure P_{sat} with an MAE of 0.298 (or 2.0% of the range of pressure values). The lowest accuracy is obtained for $K_{\mathrm{W/G}}$ with an MAE of 0.428. However, the range for partitioning coefficient $K_{\mathrm{W/G}}$ is also the largest, as seen in Figure 5, so this amounts to only 2.7% of the entire range of values. Our best machine learning MAEs are of the order of the COSMOtherm prediction accuracy, which lies at around a few tenths of log values (Stenzel et al., 2014; Schröder et al., 2016; van der Spoel et al., 2019).

Figure 6 shows the results for the best-performing descriptors MBTR and TopFP in more detail. The scatter plots illustrate how well the KRR predictions match the reference values. The match is further quantified by R^2 values. For all three target values, the predictions hug the diagonal quite closely and we observe only a few outliers that are further away from the diagonal. The predictions of partitioning coefficient $K_{\mathrm{WIOM/G}}$ are most accurate. This is expected because the MAE in Table 2 is lowest for this property. The largest scattered is observed for partitioning coefficient $K_{\mathrm{W/G}}$ which had the highest MAE in Table 2.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

Table 2. The average mean average errors (MAE) and the standard deviations for all the descriptors and target properties (equilibrium partitioning coefficients $K_{\mathrm{WIOM/G}}$, $K_{\mathrm{W/G}}$ and saturation vapour pressure P_{sat}) with the largest possible training size of 3000.

	$K_{\mathrm{WIOM/G}}$		$K_{\mathrm{W/G}}$		P_{sat}	
Descriptor	MAE	Δ	MAE	Δ	MAE log(kPa)	Δlog(kPa)
CM	0.470	± 0.020	0.787	± 0.028	0.530	± 0.016
MBTR	0.278	± 0.013	0.427	± 0.015	0.298	± 0.016
MACCS	0.412	± 0.020	0.522	± 0.020	0.431	± 0.014
Morgan	0.396	± 0.026	0.552	± 0.014	0.413	± 0.022
TopFP	0.292	± 0.014	0.451	± 0.021	0.310	± 0.014

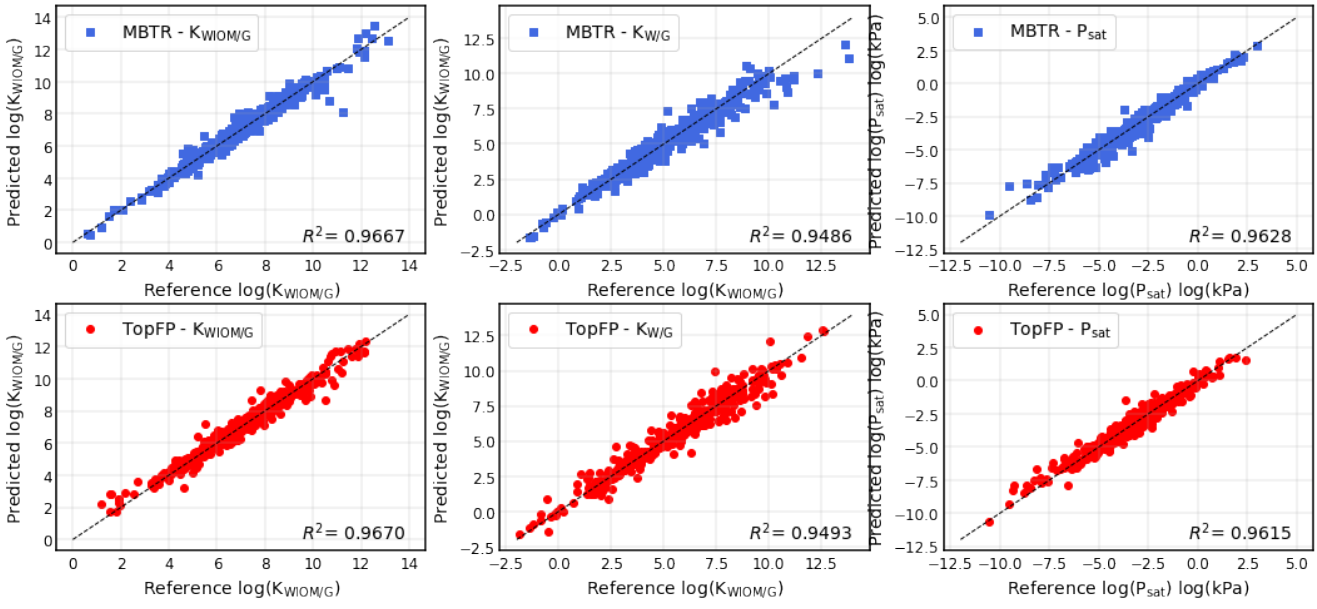

Figure 6. The scatter plots for predictions for partitioning coefficients of a molecule between a water-insoluble organic matter and gas phase $K_{\mathrm{WIOM/G}}$, water and gas phase $K_{\mathrm{W/G}}$ and the saturation vapour pressure P_{sat} for the test set of 414 molecules using MBTR (top) and TopFP (bottom). The prediction with the lowest mean average error was chosen for each scatter plot.

4 Predictions

In the previous section we showed that our KRR model trained on the Wang *et al.* dataset produces low prediction errors for molecular partitioning coefficients and can now be employed as a fast predictor. When shown further molecular structures, it

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

can make instant predictions for the molecular properties of interest. We demonstrate this application potential on an example dataset generated to imitate organic molecules typically found in the atmosphere.

Atmospheric oxidation reaction mechanisms can be generally classified into two main types: fragmentation and functionalization. For SOA formation, functionalization is more relevant, as it leads to products with intact carbon backbones and added polar (and volatility-lowering) functional groups. Many of the most interesting molecules from a SOA-forming point of view, e.g. monoterpenes, have around 10 carbon atoms. These compounds simultaneously have high enough emissions or concentrations to produce appreciable amounts of condensable products, while being large enough for those products to have low volatility.

We thus generate a dataset of molecules with a backbone of ten carbon (C10) atoms. For simplicity, we use a linear alkane chain. In analogy with Wang's dataset, we then decorate this backbone with 0 to 6 functional groups at different locations. We limit ourselves to the typical groups formed in "functionalizing" oxidation of VOC by both of the main day-time oxidants OH and O_3: carboxyl(-COOH), carbonyl (=O) and hydroxyl (-OH) (Seinfeld and Pandis, 2016). The (-COOH) group can only be added to the ends of the C10 molecule, while (=O) and (-OH) can be added to any carbon atom in the chain. We then generate all possible combinations combinatorially and filter out duplicates resulting from symmetric combinations of functional groups. In total we obtain 35,383 unique molecules. Example molecules are depicted in Figure 9.

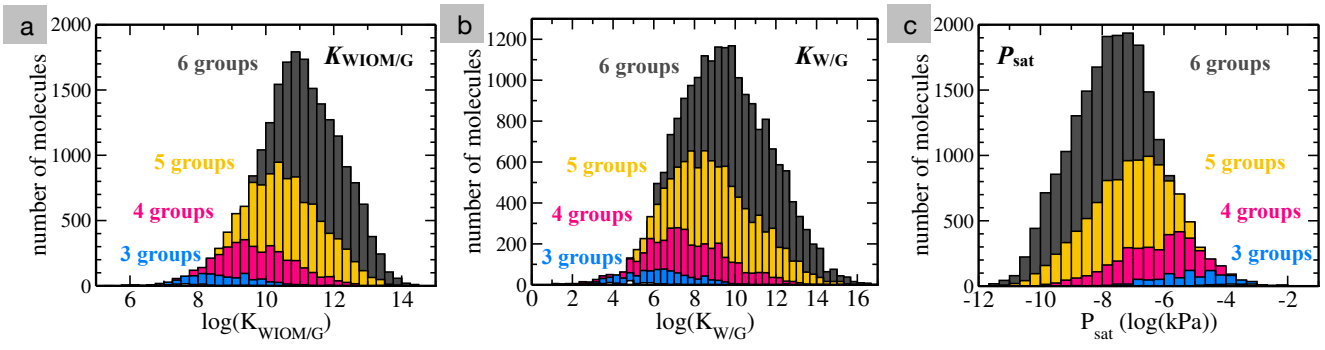

Figure 7. Histograms of C10 TopFP-KRR predictions for a) $K_{\mathrm{WIOM/G}}$,b) $K_{W/G}$ and c) P_{sat}. The histograms are divided into different numbers of functional groups. Molecules with 2 or fewer functional groups have been omitted from these histograms, because their total number is very low in the C10 dataset.

For each of the 35,383 molecules we generated a SMILES string that serves as input for the TopFP fingerprint. We did not relax the geometry of the molecules with force fields or density-functional theory. We then predicted P_{sat}, $K_{\mathrm{WIOM/G}}$ and $K_{\mathrm{W/G}}$ with the TopFP-KRR model. We chose TopFP as descriptor, because its accuracy is close to that of the best performing MBTR KRR model, but significantly cheaper to evaluate.

Figures 7 and 8 show the predictions of our TopFP-KRR model for the C10 dataset. For comparison with Wang's dataset, we broke the histograms and analysis down by the number of functional groups. For a given number of functional groups, the partitioning coefficients for our C10 dataset are somewhat higher, and the saturation vapor pressures correspondingly somewhat

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

lower, than in Wang's dataset. This follows from the fact that our C10 molecules (with between 10 and 18 non-hydrogen atoms since the largest of our molecules contain two carboxylic acid and four ketone and/or hydroxyl groups) are on average larger than those contained in Wang's dataset (Figure 2). However, as seen from Figure 8, the averages of all three quantities (for a given number of functional groups) are not substantially different, illustrating the similarity of both datasets. A certain degree of similarity is required to ensure predictive power, since machine learning models do not extrapolate well to data that lies outside the training range.

The variation in the studied parameters is larger in Wang's dataset for molecules with 4 or less functional groups or less, but similar or smaller for molecules with 5 or 6 functional groups. This is likely due to Wang's dataset containing relatively few compounds with more than four functional groups. The variation in the studied parameters (for each number of functional groups) predicted for the C10 dataset is in line with the individual group contributions predicted, based on fits to experimental data, for example by the SIMPOL model (Pankow and Asher, 2008) for saturation vapor pressures. According to SIMPOL, a carboxylic acid group decreases the saturation vapor pressure at room temperature by almost factor of 4000, while a ketone group reduces it by less than a factor of 9. Accordingly, if interactions between functional groups are ignored, for example a dicarboxylic acid should have a saturation vapor pressure more than 100 000 times lower than a diketone with the same carbon backbone. This is remarkably consistent with figure Figure 8, where the variation of saturation vapor pressures for compounds with two functional groups in our C10 dataset is slightly more than 5 orders of magnitude.

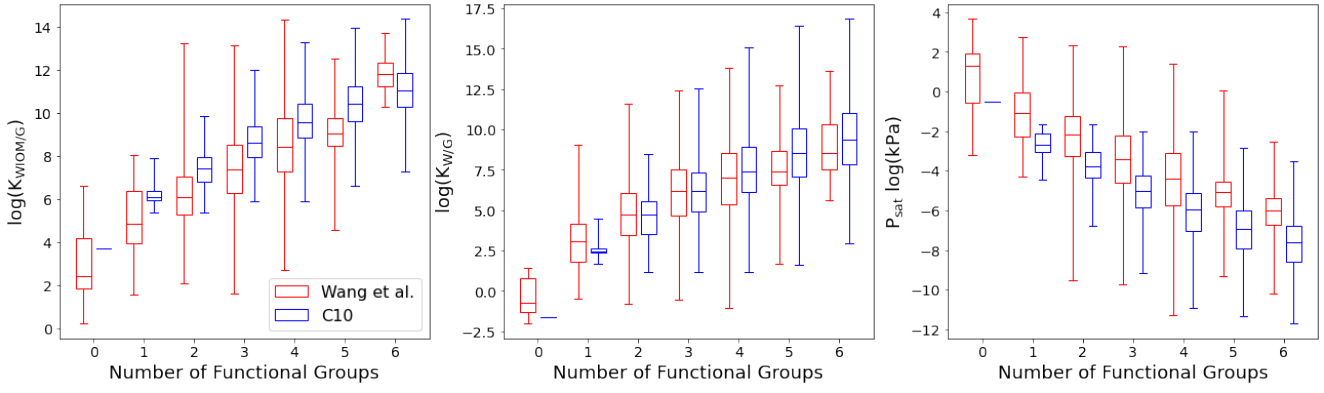

Figure 8. Box plot comparing C10 (in blue) with Wang's dataset (in red) for $K_{\mathrm{WIOM/G}}$, $K_{\mathrm{W/G}}$ and P_{sat} for different numbers of functional groups. Shown are the minimum, maximum, median, first and third quartile.

Figure 7 illustrates that the saturation vapour pressure P_{sat} decreases with increasing number of functional groups as expected, whereas $K_{\mathrm{WIOM/G}}$ and $K_{\mathrm{W/G}}$ increase. This is consistent with Wang's dataset as shown in Fig. 8, where we compare averages between the two datasets. The magnitude of the decrease (increase) amounts to approximately 1 or 2 orders of magnitude per functional group and is, again, consistent with existing structure-activity relationships based on experimental data (e.g. Pankow and Asher (2008); Compernolle et al. (2011); Nannoolal et al. (2008)).

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

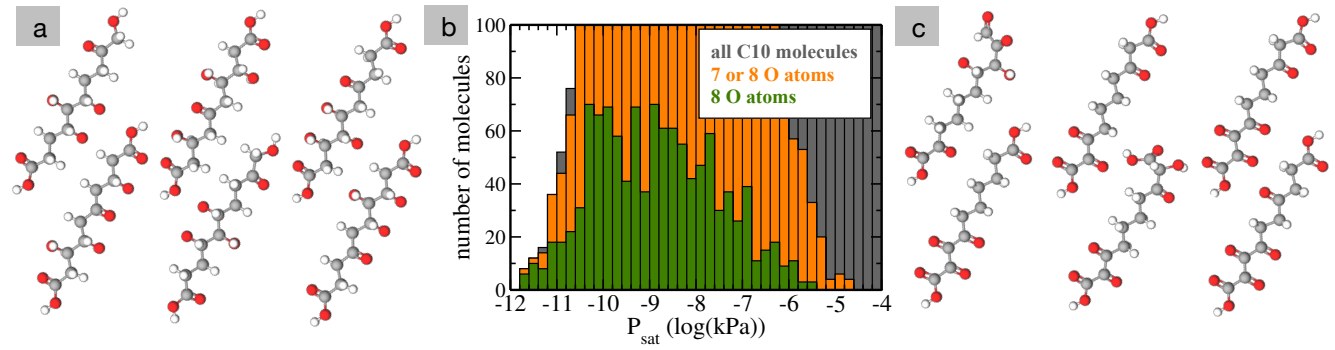

Figure 9. a) Atomic structure of the 6 molecules with the lowest predicted saturation vapour pressure P_{sat}; b) P_{sat} histograms for molecules containing 7 or 8 O atoms (orange) or only 8 O atoms (green). For reference, the histogram of all molecules (grey) is also shown. c) Atomic structure of the 6 molecules with 7 and 8 O atoms and the highest saturation vapour pressure P_{sat}.

The region of low P_{sat} is most relevant for atmospheric SOA formation. Figure 9b shows histograms of only molecules with 7 or 8 oxygen atoms. These are compared to the full dataset. Since the "8 O atom set" is a subset of the "7 or 8 O atoms" set, which in turn is a subset of "all molecules" the lengths of the bars in a given bin reflect the percentages of molecules with 7 or 8 O atoms. We observe that below 10^{-10} kPa, almost all C10 molecules contain 7 or 8 O atoms, as there is little grey visible in that part of the histogram. In the context of atmospheric chemistry, the least-volatile fraction of our C10 dataset corresponds to LVOC ("low volatility organic compounds"), which are capable of condensing onto small aerosol particles, but not actually forming them. Our results are thus in qualitative agreement with recent experimental results by (Peräkylä et al., 2020), who concluded that the highly oxidized C_{10} products of α-pinene oxidation are mostly LVOC. However, we note that the compounds measure by Peräkylä et al. are likely to contain functional groups not included in our C10 dataset, as well as structural features such as branching and rings. Figure 9a and Figure 9c show the molecular structures of the lowest-volatility compounds, as well as the highest-volatility compounds with 7 or 8 O atoms, respectively. (Note that the latter set inevitably contains at least one carboxylic acid group, as we have restricted the number of functional groups to six or less, and only the acid groups contain two oxygen atoms.) Comparing the two sets, we can see that the lowest-volatility compounds contain more hydroxyl groups, and less ketone groups, while the highest-volatility compounds with 7 or 8 oxygen atoms contain almost no hydroxyl groups. This is expected - for example according to the SIMPOL model (Pankow and Asher, 2008), a hydroxyl group lowers the saturation vapor pressure by over a factor of 100 at 298 K, while the effect of a ketone group is, as previously noted, less than a factor of 9. However, even the lowest-volatility compounds (Figure 9a) contain a few ketone groups, such that the number of hydrogen-bond donor and acceptor groups are roughly similar. This result demonstrates that unlike the simplest group-contribution models (which would invariably predict that the lowest-volatility compounds in our C10 dataset should be the tetrahydroxydicarboxylic acids), both the original COSMOTherm predictions, and the machine-learning model based on them, are capable of accounting for hydrogen-bonding interactions between functional groups.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

5 Conclusions

In this study, we set out to evaluate the potential of the KRR machine learning method to map molecular structures to its atmospheric partitioning behaviour, and establish which molecular descriptor has the best predictive capability.

KRR is a relatively simple kernel-based machine-learning technique that is straightforward to implement and fast to train. Given model simplicity, the quality of learning depends strongly on information content of the molecular descriptor. More specifically, it hinges on how well each format encapsulates the structural features relevant to the atmospheric behaviour. The exhaustive approach of MBTR descriptor to documenting molecular features has led to very good predictive accuracy in machine learning of molecular properties (Stuke et al., 2019; Langer et al., 2020; Rossi and Cumby, 2020; Himanen et al., 2020) and this work is no exception. The lightweight CM descriptor does not perform nearly as well, but these two representations from physical sciences provide us with an upper and lower limit on predictive accuracy.

Descriptors from cheminformatics that were developed specifically for molecules have variable performance. Between them, the topological fingerprint leads to best learning quality that approaches MBTR accuracy in the limit of larger training set sizes. This is a notable finding, not least because the relatively small TopFP data structures in comparison to MBTR reduce the computational time and memory required for machine learning. MBTR encoding requires knowledge of the 3-dimensional molecular structure, which raises the issue of conformer search. It is unclear which molecular conformers are relevant for atmospheric condensation behaviour, and COSMOterm calculations on different conformers can produce values that are orders of magnitude apart. TopFP requires only connectivity information and can be built from SMILES strings, eliminating any conformer considerations (albeit at the cost of possibly losing some information on e.g. intramolecular hydrogen bonds). All this makes TopFP the most promising descriptor for future machine learning studies in atmospheric science that we have identified in this work.

Our results show that KRR can be used to train a model to predict COSMOtherm saturation vapor pressures, with error margins smaller than those of the original COSMOtherm predictions. In the future, we propose to extend our training set to encompass especially atmospheric autoxidation products (Bianchi et al., 2019), which are not included in existing saturation vapour pressure datasets, and for which existing prediction methods are highly uncertain. While COSMOtherm predictions for such molecules also have large uncertainties, a fast and efficient "COSMOtherm - level" KRR predictor would still be immensely useful, for example in evaluating whether a given compound is likely to have extremely low volatility, or not. Experimental volatility data for such compounds is also gradually becoming available, either through indirect inference methods such as Peräkylä et al. (2020), or for example from thermal desorption measurements (Li et al., 2020). These can then be used to constrain and anchor the model, and ultimately yield also quantitatively reliable volatility predictions.

Code and data availability. The Wang dataset (Wang et al., 2017) and the novel C10 dataset of atmosperic molecules (Atmospheric C10 dataset, Zenodo , 2020) are freely available online. The KRR code employed in this study can be found on Gitlab (KRR for Atmospheric molecules, Gitlab , 2020).

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

Appendix A: Many-body tensor representation

In this appendix we provide the mathematical structure of the MBTR as it is implemented in the DScribe library Himanen et al. (2020). The many-body levels in the MBTR are denoted k. For $k = 1, 2, 3$, geometry functions encode the different features: $g_1(Z_l) = Z_l$ (atomic number), $g_2(\boldsymbol{R}_l, \boldsymbol{R}_m) = |\boldsymbol{R}_l - \boldsymbol{R}_m|$ (distance) or $g_2(\boldsymbol{R}_l, \boldsymbol{R}_m) = \frac{1}{|\boldsymbol{R}_l - \boldsymbol{R}_m|}$ (inverse distance), and $g_3(\boldsymbol{R}_l, \boldsymbol{R}_m, \boldsymbol{R}_n) = \cos(\angle(\boldsymbol{R}_l - \boldsymbol{R}_m, \boldsymbol{R}_n - \boldsymbol{R}_m))$ (cosine of angle).

The scalar values returned by the geometry functions g_k are Gaussian broadened into continuous representations $\mathcal{D}_k$:

$$\mathcal{D}_1^l(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x - g_1(Z_l))^2}{2\sigma_1^2}} \tag{A1}$$

$$\mathcal{D}_2^{l,m}(x) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x - g_2(\boldsymbol{R}_l, \boldsymbol{R}_m))^2}{2\sigma_2^2}} \tag{A2}$$

$$\mathcal{D}_3^{l,m,n}(x) = \frac{1}{\sigma_3\sqrt{2\pi}} e^{-\frac{(x - g_3(\boldsymbol{R}_l, \boldsymbol{R}_m), \boldsymbol{R}_n)^2}{2\sigma_3^2}}. \tag{A3}$$

The σ_k's are the feature widths for the different k-levels and x runs over a predefined range $[x^k_{\text{min}}, x^k_{\text{max}}]$ of possible values for the geometry functions g_k.

Finally, a weighted sum of distributions $\mathcal{D}_k$ is generated for each possible combination of chemical elements present in the dataset

$$\text{MBTR}_1^{Z_1}(x) = \sum_l^{|Z_1|} w_1^l \mathcal{D}_1^l(x) \tag{A4}$$

$$\text{MBTR}_2^{Z_1,Z_2}(x) = \sum_l^{|Z_1|} \sum_m^{|Z_2|} w_2^{l,m} \mathcal{D}_2^{l,m}(x) \tag{A5}$$

$$\text{MBTR}_3^{Z_1,Z_2,Z_3}(x) = \sum_l^{|Z_1|} \sum_m^{|Z_2|} \sum_n^{|Z_3|} w_3^{l,m,n} \mathcal{D}_3^{l,m,n}(x). \tag{A6}$$

The sums for l, m, and n run over all atoms with atomic numbers Z_1, Z_2 and Z_3. w_k are weighting functions that balance the relative importance of different k-terms and/or limit the range of inter-atomic interactions. For $k = 1$, usually no weighting is used ($w_1^l = 1$). For $k = 2$ and $k = 3$ the following exponential decay functions are implemented in `DScribe`

$$w_2^{l,m} = e^{-s_k|\boldsymbol{R}_l - \boldsymbol{R}_m|} \tag{A7}$$

$$w_3^{l,m,n} = e^{-s_k(|\boldsymbol{R}_l - \boldsymbol{R}_m| + |\boldsymbol{R}_m - \boldsymbol{R}_n| + |\boldsymbol{R}_l - \boldsymbol{R}_n|)} \tag{A8}$$

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

Table A1. Optimal KRR hyperparameter α values obtained by cross-validation as a function of descriptor type and training set size. The procedure was repeated 10 times with re-shuffled data. Average values ($\bar{\alpha}$) were used in further KRR models. We also report the statistical standard deviation $\Delta\alpha$.

Descriptor	Training Set Size	$K_{WIOM/G}$		$K_{W/G}$		P_{sat}	
		$\bar{\alpha}$	$\Delta\alpha$	$\bar{\alpha}$	$\Delta\alpha$	$\bar{\alpha}$	$\Delta\alpha$
CM	500	0.00E+00	4.50E-03	9.10E-03	2.85E-03	1.00E-02	0.00E+00
	1000	8.20E-03	3.79E-03	5.50E-03	4.74E-03	1.00E-02	0.00E+00
	1500	5.50E-03	4.74E-03	3.70E-03	4.35E-03	5.50E-03	4.74E-03
	2000	3.70E-03	4.35E-03	1.81E-03	2.89E-03	5.50E-03	4.74E-03
	2500	2.80E-03	3.79E-03	1.00E-03	0.00E+00	1.90E-03	2.85E-03
	3000	1.90E-03	2.85E-03	1.00E-03	0.00E+00	1.90E-03	2.85E-03
MBTR	500	5.30E-05	4.97E-05	7.30E-05	4.35E-05	1.44E-04	3.04E-04
	1000	8.02E-05	4.17E-05	1.00E-04	0.00E+00	1.81E-04	2.89E-04
	1500	1.72E-04	2.93E-04	2.62E-04	3.91E-04	2.71E-04	3.85E-04
	2000	2.53E-04	3.96E-04	3.16E-04	4.73E-04	5.32E-04	4.94E-04
	2500	6.04E-04	5.11E-04	7.03E-04	4.78E-04	9.01E-04	3.13E-04
	3000	7.03E-04	4.78E-04	5.05E-04	5.22E-04	1.00E-03	0.00E+00
TopFP	500	5.50E-03	4.74E-03	3.70E-03	4.35E-03	9.10E-03	2.85E-03
	1000	8.20E-03	3.79E-03	1.90E-03	2.85E-03	7.30E-03	4.35E-03
	1500	8.20E-03	3.79E-03	3.70E-03	4.35E-03	9.10E-03	2.85E-03
	2000	1.00E-02	0.00E+00	1.00E-03	0.00E+00	9.10E-03	2.85E-03
	2500	7.30E-03	4.35E-03	1.00E-03	0.00E+00	8.20E-03	3.79E-03
	3000	7.30E-03	4.35E-03	1.00E-03	0.00E+00	9.10E-03	2.85E-03
MACCS	500	1.00E-02	0.00E+00	3.10E-02	4.65E-02	1.00E-02	0.00E+00
	1000	1.00E-02	0.00E+00	9.10E-02	2.83E-02	1.00E-02	0.00E+00
	1500	1.00E-02	0.00E+00	1.00E-01	0.00E+00	1.00E-02	0.00E+00
	2000	8.20E-03	2.84E-03	1.00E-01	0.00E+00	5.50E-02	4.47E-02
	2500	2.80E-03	3.74E-03	1.00E-01	0.00E+00	9.01E-02	3.11E-02
	3000	1.00E-03	0.00E+00	1.00E-01	0.00E+00	9.01E-02	3.11E-02
Morgan	500	4.00E-04	4.72E-04	7.10E-03	4.57E-03	5.10E-03	4.86E-03
	1000	3.00E-03	4.72E-03	1.00E-02	0.00E+00	9.10E-03	2.83E-03
	1500	1.00E-02	0.00E+00	1.90E-02	2.83E-02	1.00E-02	0.00E+00
	2000	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00
	2500	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00
	3000	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

Table A2. Optimal KRR hyperparameter γ values obtained by cross-validation as a function of descriptor type and training set size. The procedure was repeated 10 times with re-shuffled data. Average values ($\bar{\gamma}$) were used in further KRR models. We also report the statistical standard deviation $\Delta\gamma$.

Descriptor	Training Set Size	$K_{WIOM/G}$		$K_{W/G}$		P_{sat}	
		$\bar{\gamma}$	$\Delta\gamma$	$\bar{\gamma}$	$\Delta\gamma$	$\bar{\gamma}$	$\Delta\gamma$
CM	500	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	1000	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	1500	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	2000	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	2500	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	3000	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
MBTR	500	5.30E-05	5.30E-05	7.30E-05	4.35E-05	5.40E-05	4.86E-05
	1000	8.20E-05	8.20E-05	1.00E-04	0.00E+00	1.81E-04	2.89E-04
	1500	1.90E-04	1.90E-04	2.80E-04	3.79E-04	2.80E-04	3.79E-04
	2000	2.80E-04	2.80E-04	3.70E-04	4.35E-04	5.50E-04	4.74E-04
	2500	6.40E-04	6.40E-04	7.30E-04	4.35E-04	9.10E-04	2.85E-04
	3000	7.30E-04	7.30E-04	5.50E-04	4.74E-04	1.00E-03	0.00E+00
TopFP	500	1.00E-04	0.00E+00	9.10E-05	2.85E-05	1.00E-04	0.00E+00
	1000	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	1500	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	2000	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	2500	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
	3000	1.00E-04	0.00E+00	1.00E-04	0.00E+00	1.00E-04	0.00E+00
MACCS	500	1.00E-02	0.00E+00	9.10E-02	2.83E-02	1.00E-02	0.00E+00
	1000	1.00E-02	0.00E+00	9.10E-02	2.83E-02	1.00E-02	0.00E+00
	1500	1.00E-02	0.00E+00	1.00E-01	0.00E+00	1.00E-02	0.00E+00
	2000	1.00E-02	0.00E+00	1.00E-01	0.00E+00	5.50E-02	4.47E-02
	2500	1.00E-02	0.00E+00	1.00E-01	0.00E+00	9.10E-02	2.83E-02
	3000	1.00E-02	0.00E+00	1.00E-01	0.00E+00	9.10E-02	2.83E-02
Morgan	500	4.00E-04	4.72E-04	9.00E-03	3.14E-03	5.10E-03	4.86E-03
	1000	3.00E-03	4.72E-03	1.00E-02	0.00E+00	9.01E-03	3.11E-03
	1500	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00
	2000	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00
	2500	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00
	3000	1.00E-02	0.00E+00	1.00E-02	0.00E+00	1.00E-02	0.00E+00

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021


Atmospheric Chemistry and Physics Discussions

The parameter s_k effectively tunes the cutoff distance. The functions $\mathrm{MBTR}_k(x)$ are then discretized with n_k many points in the respective intervals $[x^k_{\mathrm{min}}, x^k_{\mathrm{max}}]$.

Author contributions. EL performed all computational work. ML advised the computations. PR, HV and TK conceived the study. All authors participated in drafting the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Academy of Finland (Project numbers 315600 and 316601) and through their Flagship programme: Finnish Center for Artificial Intelligence FCAI. This work was further supported by the European Research Council project 692891-DAMOCLES, by COST (European Cooperation in Science and Technology) Action 18234 and by the University of Helsinki Faculty of Science ATMATH project. We thank CSC, the Finnish IT Center for Science and Aalto Science IT for computational resources.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Atmospheric Chemistry and Physics Discussions

References

Arp, H. P. H. and Goss, K.-U.: Ambient Gas/Particle Partitioning. 3. Estimating Partition Coefficients of Apolar, Polar, and Ionizable Organic Compounds by Their Molecular Structure, Environ. Sci. Technol., 43, 1923–1929, 2009.

Atmospheric C10 dataset, Zenodo (2020): Https://doi.org/10.5281/zenodo.4291795.

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., and Anderson, D.: Viewing Forced Climate Patterns Through an AI Lens, Geophys. Res. Lett., 46, 13 389–13 398, 2019.

Bartók, A. P., De, S., Poelking, C., Bernstein, N., Kermode, J. R., Csányi, G., and Ceriotti, M.: Machine learning unifies the modeling of materials and molecules, Sci. Adv., 3, e1701 816, 2017.

Bianchi, F., Kurtén, T., Riva, M., Mohr, C., Rissanen, M. P., Roldin, P., Berndt, T., Crounse, J. D., Wennberg, P. O., Mentel, T. F., Wildt, J., Junninen, H., Jokinen, T., Kulmala, M., Worsnop, D. R., Thornton, J. A., Donahue, N., Kjaergaard, H. G., and Ehn, M.: Highly Oxygenated Organic Molecules (HOM) from Gas-Phase Autoxidation Involving Peroxy Radicals: A Key Contributor to Atmospheric Aerosol, Chem. Rev., 119, 3472–3509, 2019.

Cervone, G., P, F., Ezber, Y., and Boybeyi, Z.: Risk assessment of atmospheric emissions using machine learning, Nat. Hazards Earth Syst. Sci., 8, 991–1000, 2008.

Coley, C. W., Eyke, N. S., and Jensen, K. F.: Autonomous discovery in the chemical sciences part II: Outlook, Angew. Chem. Int. Ed., n/a.

Compernolle, S., Ceulemans, K., and Müller, J.-F.: EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions, Atmospheric Chem. Phys., 11, 9431–9450, 2011.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G.: Reoptimization of MDL Keys for Use in Drug Discovery, J. Chem. Inform. Comput. Sci., 42, 1273–1280, 2002.

Eckert, F. and Klamt, A.: Fast solvent screening via quantum chemistry: COSMO-RS approach, AIChE J., 48, 369–385, 2002.

Elm, J., Kubečka, J., Besel, V., Jääskeläinen, M. J., Halonen, R., Kurtén, T., and Vehkamäki, H.: Modeling the formation and growth of atmospheric molecular clusters: A review, J. Aerosol Sci., 149, 105 621, 2020.

Faber, F., Lindmaa, A., Lilienfeld, O. A. v., and Armiento, R.: Crystal structure representations for machine learning models of formation energies, Int. J. Quantum Chem., 115, 1094–1101, 2015.

Friedman, J., Hastie, T., and Tibshirani, R.: The elements of statistical learning, vol. 1, Springer series in statistics New York, 2001.

Ghosh, K., Stuke, A., Todorović, M., Jørgensen, P. B., Schmidt, M. N., Vehtari, A., and Rinke, P.: Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra, Adv. Sci., 6, 1801 367, 2019.

Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J., and Sutton, C.: Machine learning for heterogeneous catalyst design and discovery, AIChE Journal, 64, 2311–2323, 2018.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T. C.-C., Markopoulos, G., Jeon, S., Kang, H., Miyazaki, H., Numata, M., Kim, S., Huang, W., Hong, S. I., Baldo, M. A., Adams, R. P., and Aspuru-Guzik, A.: Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach., Nat. Mater., 15, 1120–7, 2016.

Goss, K.-U.: The Air/Surface Adsorption Equilibrium of Organic Compounds Under Ambient Conditions, Crit. Rev. Env. Sci. Tec., 34, 339–389, 2004.

Goss, K.-U.: Prediction of the temperature dependency of Henry's law constant using poly-parameter linear free energy relationships, Chemosphere, 64, 1369 – 1374, 2006.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Goss, K.-U. and Schwarzenbach, R. P.: Linear Free Energy Relationships Used To Evaluate Equilibrium Partitioning of Organic Compounds, Environ. Sci. Technol., 35, 1–9, 2001.

Gu, G. H., Noh, J., Kim, I., and Jung, Y.: Machine learning for renewable energy materials, J. Mater. Chem. A, pp. 17 096–17 117, 2019.

Hilal, S. H., Ayyampalayam, S. N., and Carreira, L. A.: Air-Liquid Partition Coefficient for a Diverse Set of Organic Compounds: Henry's Law Constant in Water and Hexadecane, Environ. Sci. Technol., 42, 9231–9236, 2008.

Himanen, L., Geurts, A., Foster, A. S., and Rinke, P.: Data-Driven Materials Science: Status, Challenges, and Perspectives, Adv. Sci., 6, 1900 808, 2019.

Himanen, L., Jäger, M. O. J., Morooka, E. V., Canova, F. F., Ranawat, Y. S., Gao, D. Z., Rinke, P., and Foster, A. S.: DScribe: Library of descriptors for machine learning in materials science, Comp. Phys. Commun., 247, 106 949, 2020.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., and Yang, H.: Machine learning and artificial intelligence to aid climate change research and preparedness, Environ. Res. Lett., 14, 124 007, 2019.

Huo, H. and Rupp, M.: Unified Representation for Machine Learning of Molecules and Crystals, arXiv:1704.06439, 2017.

James, C., Weininger, D., and Delany, J.: Daylight Theory Manual. Daylight Chemical Information Systems, Inc., Irvine, CA, 1995.

Jenkin, M. E., Saunders, S. M., and Pilling, M. J.: The tropospheric degradation of volatile organic compounds: a protocol for mechanism development, Atmos. Environ., 31, 81 – 104, 1997.

Jensen, K. F., Coley, C. W., and Eyke, N. S.: Autonomous discovery in the chemical sciences part I: Progress, Angew. Chem. Int. Ed., n/a.

Kalberer, M., Paulsen, D., Sax, M., Steinbacher, M., Dommen, J., Prevot, A. S. H., Fisseha, R., Weingartner, E., Frankevich, V., Zenobi, R., and Baltensperger, U.: Identification of Polymers as Major Components of Atmospheric Organic Aerosols, Science, 303, 1659–1662, 2004.

Klamt, A. and Eckert, F.: COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, Fluid Phase Equilib., 172, 43 – 72, 2000.

Klamt, A. and Eckert, F.: Erratum to "COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids" [Fluid Phase Equilib. 172 (2000) 43–72], Fluid Phase Equilib., 205, 357, 2003.

KRR for Atmospheric molecules, Gitlab (2020): https://gitlab.com/cest-group/krr-and-atmospheric-molecules.

Landrum, G. et al.: RDKit: Open-source cheminformatics, 2006.

Langer, M. F., Goeßmann, A., and Rupp, M.: Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning, arXiv:2003.12081, 2020.

Li, Z., D'Ambro, E. L., Schobesberger, S., Gaston, C. J., Lopez-Hilfiker, F. D., Liu, J., Shilling, J. E., Thornton, J. A., and Cappa, C. D.: A robust clustering algorithm for analysis of composition-dependent organic aerosol thermal desorption measurements, Atmospheric Chem. Phys., 20, 2489–2512, 2020.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V.: Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, J. Chem. Inf. Model, 55, 263–274, 2015.

Masuda, R., Iwabuchi, H., Schmidt, K. S., Damiani, A., and Kudo, R.: Retrieval of Cloud Optical Thickness from Sky-View Camera Images using a Deep Convolutional Neural Network based on Three-Dimensional Radiative Transfer, Remote Sens., 11, 1962, 2019.

Meyer, B., Sawatlon, B., Heinen, S., von Lilienfeld, O. A., and Corminboeuf, C.: Machine learning meets volcano plots: computational discovery of cross-coupling catalysts, Chem. Sci., 9, 7069–7077, 2018.

Morgan, H. L.: The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service., J. Chem. Doc., 5, 107–113, 1965.

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

Müller, T., Kusne, A. G., and Ramprasad, R.: Machine Learning in Materials Science, chap. 4, pp. 186–273, John Wiley & Sons, Ltd, Hoboken, New Jersey, USA, 2016.

Nannoolal, Y., Rarey, J., and Ramjugernath, D.: Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilib., 269, 117 – 133, 2008.

Nourani, V., Uzelaltinbulat, S., Sadikoglu, F., and Behfar, N.: Artificial Intelligence Based Ensemble Modeling for Multi-Station Prediction of Precipitation, Atmosphere, 10, 80, 2019.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R.: Open Babel: An open chemical toolbox, J. Cheminform., 3, 33, 2011.

Pankow, J. F. and Asher, W. E.: SIMPOL.1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds, Atmospheric Chem. Phys., 8, 2773–2796, 2008.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Peräkylä, O., Riva, M., Heikkinen, L., Quéléver, L., Roldin, P., and Ehn, M.: Experimental investigation into the volatilities of highly oxygenated organic molecules (HOMs), Atmospheric Chem. Phys., 20, 649–669, 2020.

Pyzer-Knapp, E. O., Li, K., and Aspuru-Guzik, A.: Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery, Adv. Funct. Mater, 25, 6495–6502.

Raventos-Duran, T., Camredon, M., Valorso, R., Mouchel-Vallon, C., and Aumont, B.: Structure-activity relationships to estimate the effective Henry's law constants of organics of atmospheric interest, Atmos. Chem. Phys., 10, 7643–7654, 2010.

Rogers, D. and Hahn, M.: Extended-connectivity fingerprints, J. Chem. Inf. Model., 50, 742–754, 2010.

Rossi, K. and Cumby, J.: Representations and descriptors unifying the study of molecular and bulk systems, Int. J. Quantum Chem., 120, e26 151, 2020.

Rupp, M.: Machine learning for quantum mechanics in a nutshell, Int. J. Quantum Chem., 115, 1058–1073, 2015.

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A.: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett., 108, 058 301, 2012.

Rupp, M., von Lilienfeld, O. A., and Burke, K.: Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry, J. Chem. Phys., 148, 241 401, 2018.

Sander, R.: Compilation of Henry's law constants (version 4.0) for water as solvent, Atmos. Chem. Phys., 15, 4399–4981, 2015.

Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J.: Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds, Atmos. Chem. Phys., 3, 161–180, 2003.

Schmidt, J., Marques, M. R. G., Botti, S., and Marques, M. A. L.: Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater., 5, 83, 2019.

Schröder, B., Fulem, M., and M. A.R. Martins: Vapor pressure predictions of multi-functional oxygen-containing organic compounds with COSMO-RS, Atmos. Environ., 133, 135 – 144, 2016.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 3rd Edition, Wiley, 2016.

Shandiz, M. A. and Gauvin, R.: Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries, Comput. Mater. Sci., 117, 270 – 278, 2016.

Shrivastava, M., Andreae, M. O., Artaxo, P., Barbosa, H. M. J., Berg, L. K., Brito, J., Ching, J., Easter, R. C., Fan, J., Fast, J. D., Feng, Z., Fuentes, J. D., Glasius, M., Goldstein, A. H., Alves, E. G., Gomes, H., Gu, D., Guenther, A., Jathar, S. H., Kim, S., Liu, Y., Lou, S., Martin,

https://doi.org/10.5194/acp-2020-1258
Preprint. Discussion started: 26 January 2021

S. T., McNeill, V. F., Medeiros, A., de Sá, S. S., Shilling, J. E., Springston, S. R., Souza, R. A. F., Thornton, J. A., Isaacman-VanWertz, G., Yee, L. D., Ynoue, R., Zaveri, R. A., Zelenyuk, A., and Zhao, C.: Urban pollution greatly enhances formation of natural aerosols over the Amazon rainforest, Nat. Commun., 10, 1046, 2019.

Smith, J. S., Isayev, O., and Roitberg, A. E.: ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, Chem. Sci., 8, 3192–3203, 2017.

Stenzel, A., Goss, K.-U., and Endo, S.: Prediction of partition coefficients for complex environmental contaminants: Validation of COSMOtherm, ABSOLV, and SPARC, Environ. Toxicol. Chem., 33, 1537–1543, 2014.

Stuke, A., Todorović, M., Rupp, M., Kunkel, C., Ghosh, K., Himanen, L., and Rinke, P.: Chemical diversity in molecular orbital energy predictions with kernel ridge regression, J. Chem. Phys., 150, 204 121, 2019.

Stuke, A., Rinke, P., and Todorović, M.: Efficient hyperparameter tuning for kernel ridge regression with Bayesian optimization, arXiv:2004.00675, 2020.

Todorović, M., Gutmann, M. U., Corander, J., and Rinke, P.: Bayesian inference of atomistic structure in functional materials, npj Comp. Mat., 5, 35, 2019.

Toms, B. A., Kashinath, K., Prabhat, M., Mudigonda, M., and Yang, D.: Climate Science, Deep Learning, and Pattern Discovery: The Madden-Julian Oscillation as a Test Case, in: AGU Fall Meeting Abstracts, vol. 2018, pp. IN21D–0738, 2018.

Topping, D., Barley, M., Bane, M. K., Higham, N., Aumont, B., Dingle, N., and McFiggans, G.: UManSysProp v1.0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations, Geosci. Model Dev., 9, 899–914, 2016.

Valorso, R., Aumont, B., Camredon, M., Raventos-Duran, T., Mouchel-Vallon, C., Ng, N. L., Seinfeld, J. H., Lee-Taylor, J., and Madronich, S.: Explicit modelling of SOA formation from α-pinene photooxidation: sensitivity to vapour pressure estimation, Atmospheric Chem. Phys., 11, 6895–6910, 2011.

van der Spoel, D., Manzetti, S., Zhang, H., and Klamt, A.: Prediction of Partition Coefficients of Environmental Toxins Using Computational Chemistry Methods, ACS Omega, 4, 13 772–13 781, 2019.

Wang, C., Yuan, T., Wood, S. A., Goss, K.-U., Li, J., Ying, Q., and Wania, F.: Uncertain Henry's law constants compromise equilibrium partitioning calculations of atmospheric oxidation products, Atmos. Chem. Phys., 17, 7529–7540, 2017.

Ye, Q., Robinson, E. S., Ding, X., Ye, P., Sullivan, R. C., and Donahue, N. M.: Mixing of secondary organic aerosols versus relative humidity, Proc. Natl. Acad. Sci., 113, 12 649–12 654, 2016.

Zunger, A.: Inverse design in search of materials with target functionalities, Nat. Rev. Chem., 2, 0121, 2018.